Appl. No. 09/645,479
Amdt. dated January 21, 2005
Amendment/RCE Submission

PATENT

Amendments to the Claims:

This listing of claims will replace all prior versions, and listings, of claims in the application:

Listing of Claims:

1 1. (Currently Amended) A computer implemented method of identifying
2 ~~and extracting~~ desired content ~~from~~ in HTML formatted web pages, comprising the steps of:
3 selecting a model page, wherein the model page includes content data and a
4 plurality of HTML tags for formatting the content data;
5 identifying a first area of interest in the model page;
6 parsing the model page to generate a first string of symbols ~~corresponding to each~~
7 ~~of~~ for the plurality of HTML tags, the generated symbols in the first string representing only
8 HTML tags, wherein the first area of interest is identified by a first portion of the first string of
9 symbols;
10 retrieving a second web page associated with a different URL than the model
11 page;
12 parsing the second web page to generate a second string of symbols
13 ~~corresponding to each of the~~ for a plurality of HTML tags of the second web page, the generated
14 symbols in the second string representing only HTML tags; and
15 comparing the first and second symbol strings to determine whether the second
16 string includes a second portion similar to the first portion of the first string, wherein the second
17 portion corresponds to a second area of interest in the second page.

1 2. (Original) The method of claim 1, wherein the step of comparing
2 includes applying an approximate pattern matching algorithm to the first and second strings.

1 3. (Original) The method of claim 1, further comprising the step of
2 storing the first and second areas of interest in a database.

Appl. No. 09/645,479
Amdt. dated January 21, 2005
Amendment/RCE Submission

PATENT

1
2

4.   (Currently amended) The method of claim 1, further comprising the step of extracting content data in the second area of interest from the second page.

1
2

5.   (Original)   The method of claim 4, further comprising the step of applying a regular expression matching algorithm to the extracted second area of interest.

1
2

6.   (Original)   The method of claim 1, wherein the first and second areas of interest each include two or more distinct sub-areas of the respective page.

1
2
3

7.   (Original)   The method of claim 1, wherein the step of identifying a first area of interest includes the step of identifying portions of the HTML tags of the model page.

1
2

8.   (Original)   The method of claim 1, wherein the step of identifying a first area of interest is performed using a manual pointing and selecting device.

1
2
3

9.   (Original)   The method of claim 1, wherein the steps of selecting and identifying are performed manually and wherein the remaining steps are performed automatically.

1
2

10.   (Original)   The method of claim 1, wherein the second web page is retrieved from a remote website over the Internet.

1
2

11.   (Original)   The method of claim 1, wherein the HTML tags include attributes and attribute values.

1
2
3
4
5
6

12.   (Currently amended) A computer readable medium containing instructions for controlling a computer system to automatically identify and extract desired content from in a retrieved HTML formatted web page, by automatically:

    parsing the HTML code of a manually selected model web page to generate a first string of symbols corresponding to each of for a first plurality of HTML tags, the generated symbols in the first string representing only HTML tags;

7   retrieving a second web page associated with a different URL than the model web

8   page;

9   parsing the HTML code of the second web page to generate a second string of

10   symbols ~~corresponding to each of the~~ for HTML tags of the second page, the generated symbols

11   in the second string representing only HTML tags; and

12   comparing the first and second symbol strings to determine whether the second

13   page includes a second plurality of HTML tags substantially matching the first plurality of

14   HTML tags.

1   13.   (Original)   The computer readable medium of claim 12, wherein the

2   first plurality of HTML tags are identified by an operator using a pointing and selection device

3   coupled to the computer system.

1   14.   (Original)   The computer readable medium of claim 12, wherein the

2   second web page is retrieved from a remote website over the Internet.

1   15.   (Original)   The computer readable medium of claim 12, further

2   including instructions for extracting a portion of the second page corresponding to the second

3   plurality of HTML tags.

1   16.   (Original)   The computer readable medium of claim 15, wherein the

2   instructions further control the computer system to store the extracted portion of the second page

3   in a database.

1   17.   (Original)   The computer readable medium of claim 15, further

2   including instructions for controlling the computer system to apply a regular expression

3   matching algorithm to the extracted portion of the second page.

1   18.   (Original)   The computer readable medium of claim 15, wherein the

2   extracted portion of the second page includes two or more distinct sub-areas.

Appl. No. 09/645,479
Amdt. dated January 21, 2005
Amendment/RCE Submission

PATENT

.1   19.   (Original)   The computer readable medium of claim 12, wherein the

2   instructions for comparing include instructions for applying an approximate string matching

3   algorithm to the first and second strings.

1   20.   (Original)   The computer readable medium of claim 12, wherein the

2   HTML tags include attributes and attribute values.

1   21.   (Currently amended) A computer system for identifying and extracting

2   content from HTML formatted web pages, the system comprising:

3   means for retrieving web pages including <u>content data and</u> HTML tags <u>for</u>

4   <u>formatting the content data</u>, wherein a model web page is retrieved;

5   means for manually identifying a first area of interest in the model page, wherein

6   the first area of interest corresponds to a first plurality of HTML tags; and

7   a processor including:

8   means for parsing a page, wherein the parsing means parses the model page and

9   generates a first string of symbols ~~corresponding to each of~~ <u>for</u> the first plurality of HTML tags<u>,</u>

10   <u>the generated symbols in the first string representing only HTML tags</u>, and wherein the parsing

11   means thereafter parses an automatically retrieved second web page associated with a different

12   URL than the model page and generates a second string of symbols ~~corresponding to each of the~~

13   <u>for</u> HTML tags of the second web page<u>, the generated symbols in the second string representing</u>

14   <u>only HTML tags</u>;

15   means for comparing the first and second <u>symbol</u> strings to determine whether the

16   second string includes a second portion similar to the first portion of the first string, wherein the

17   second portion corresponds to a second area of interest in the second page; and

18   means for extracting <u>content data in</u> the second area of interest from the second

19   page.

Appl. No. 09/645,479
Amdt. dated January 21, 2005
Amendment/RCE Submission

PATENT

.1   22. (Currently amended) A computer implemented method of identifying

2 ~~and extracting~~ desired content ~~from~~ in web pages formatted using a markup language,

3 comprising the steps of:

4   selecting a model page, wherein the model page includes a plurality of tokens,

5 wherein tokens include HTML tag elements and content elements;

6   identifying a first area of interest in the model page;

7   parsing the model page to generate a first string of symbols ~~corresponding to each~~

8 ~~of~~ for the plurality of tokens in the model page, the generated symbols in the first string

9 representing only tag elements, wherein the first area of interest is identified by a first portion of

10 the first string of symbols;

11   retrieving a second web page associated with a different URL than the model

12 page;

13   parsing the second web page to generate a second string of symbols

14 ~~corresponding to each of the~~ for a plurality of tokens of the second web page, the generated

15 symbols in the second string representing only tag elements; and

16   comparing the first and second symbol strings to determine whether the second

17 string includes a second portion similar to the first portion of the first string, wherein the second

18 portion corresponds to a second area of interest in the second page.

1   23. (Currently amended) The method of claim 22, further comprising the step

2 of extracting content elements in the second area of interest from the second page.

1   24. (Original) The method of claim 22, wherein the markup language is

2 selected from the group consisting of HTML, XML, WML, DHTML and HDML.

1   25. (Canceled).

1   26. (Currently amended) A computer-implemented method of identifying

2 similar content in HTML formatted web pages, the method comprising:

Appl. No. 09/645,479
Amdt. dated January 21, 2005
Amendment/RCE Submission

PATENT

3   selecting a model page, wherein the model page includes <u>content data and</u> a

4   plurality of HTML tags <u>for formatting the content data</u>;

5   identifying a first area of interest in the model page;

6   generating a first string of symbols for the plurality of HTML tags associated with

7   the first area of interest<u>, the generated symbols in the first string representing only HTML tags</u>;

8   ~~each symbol corresponding to a different one of the plurality of HTML tags;~~

9   retrieving a second web page associated with a different URL than the model

10   page;

11   generating a second string of symbols for the HTML tags of the second web page<u>,</u>

12   <u>the generated symbols in the second string representing only HTML tags</u>; ~~each second symbol~~

13   ~~corresponding to a different one of the plurality of HTML tags of the second web page;~~ and

14   comparing the first and second <u>symbol</u> strings to determine whether the second

15   string includes a portion similar to the first string, wherein the portion corresponds to a second

16   area of interest in the second page.

1   27.   (Currently amended) The method of claim 26, further comprising

2   extracting <u>content data in</u> the second area of interest from the second page.

1   28.   (Previously presented) The method of claim 26, wherein identifying is

2   performed manually using a user-input device.